Projet d'analyse de données - Revue finale

Manon Tessier (GI05), Sacha Benarroch-Lelong (GI04)

SY09 - Printemps 2022, 15 juin 2022

Résumé

Ce projet d'analyse de données a pour objectif de mettre en oeuvre diverses techniques afin d'extraire autant d'informations que possible d'un jeu de données.

Pour cette analyse, nous avons choisi de travailler sur le jeu de données `Friends` (2020), disponible sur le dépôt Git du *TidyTuesday*[1].

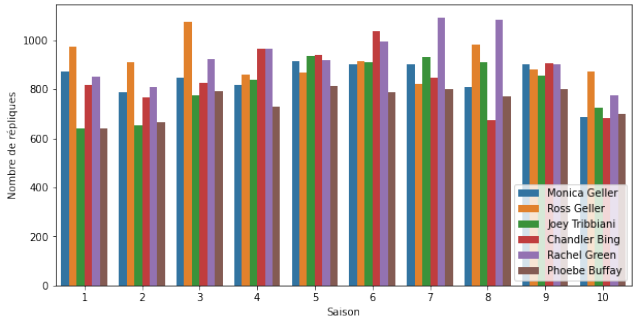


Figure 1 – Nombre de répliques des personnages principaux

1 Analyse exploratoire

1.1 Présentation des données

Le jeu de données est composé de trois fichiers CSV :
- `friends.csv` regroupe 67373 lignes de dialogue extraites du script de la série américaine *Friends* ;
- `friends_emotions.csv` associe à certaines de ces lignes de dialogue des émotions ;
- `friends_info.csv` fournit des informations sur les 236 épisodes des 10 saisons de la série.

Un rapide pré-traitement des données a été effectué, notamment des transtypages et des éliminations de doublons, ainsi qu'une jointure entre les 3 tables (jointure externe gauche pour ne perdre aucune donnée dû au caractère partiel de la table `friends_emotions.csv`).

1.2 Quantité de répliques

L'aspect le plus évident à exploiter du jeu de données est la proportion des interventions de chacun des personnages principaux, au nombre de 6.

1.2.1 Représentativité

Un préliminaire à cette analyse est de s'assurer de la représentativité relative de l'échantillon de dialogues sur toute la série. On observe pour cela le nombre de répliques par épisode. On obtient un écart interquartile $q_3 - q_1 = 41$ et un écart-type $\sigma \simeq 39.39$, pour une médiane de 283.5 répliques. Nous considérons les valeurs des mesures de dispersion suffisamment faibles au regard de la médiane pour penser que chaque épisode (et donc chaque saison) est correctement représenté(e). Rien en revanche ne nous permet d'affirmer la représentativité des données propres à chaque épisode, la méthode de sélection des répliques n'étant pas détaillée dans la documentation. Nous ferons l'hypothèse que cette sélection est aléatoire, et donc que le contenu sélectionné dans chaque épisode en est représentatif. Toutes les conclusions qui seront tirées de ce jeu de données sont soumises à cette hypothèse.

1.2.2 Répliques des personnages principaux

Les interventions de chacun des personnages principaux dans les répliques que nous étudions, sont distribuées selon ce qui est montré en figure 1. Cette figure montre que le couple Rachel/Ross prédomine dans les premières saisons, puis qu'il est rattrapé et parfois dépassé par Chandler et Monica au fil du temps. Les deux personnages le moins exploités ont des dynamiques dépendantes des saisons : Joey possède de plus en plus de répliques, alors que le nombre varie plus aléatoirement pour Phoebe.

Outre l'effet de la saison, nous nous sommes intéressés à celui de l'auteur de chaque épisode sur la quantité de dialogue allouée à chaque personnage. Une analyse

1. https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-09-08/readme.md

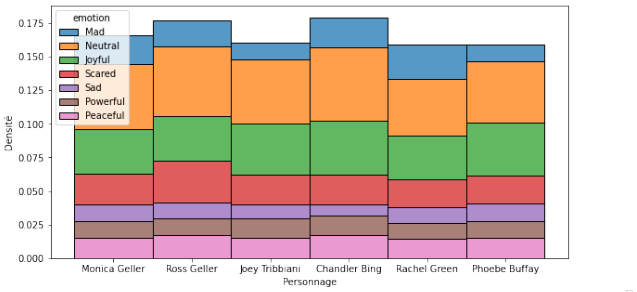


Figure 2 – Répartition des émotions dans les répliques des personnages principaux

de la variance est réalisée pour chaque personnage sur les quantités de dialogues dans chaque épisode, regroupés par auteur (pour les 7 auteurs ayant écrit plus de 10 épisodes). Des tests de normalité et d'homoscédasticité sont effectués au préalable (tests de Shapiro-Wilk et de Bartlett). Certains auteurs ont dû être éliminés du jeu de données pour certains personnages après rejet de l'hypothèse de normalité des quantités de dialogue dans leurs épisodes. Leur nombre, ainsi que les *p-value* associées aux ANOVA effectuées sur les données restantes, sont reportées dans la table 1.

Table 1 – Résultat de l'ANOVA sur les quantités de dialogue par auteur.

Personnage	Auteurs éliminés	α^* associé
Monica	1	0.8051
Ross	1	0.0298
Joey	1	0.0001
Chandler	1	0.2080
Rachel	1	0.0147
Phoebe	1	0.9192

À un niveau de confiance de 5%, l'hypothèse d'égalité des moyennes peut donc être rejetée dans les cas de Ross, Joey et Rachel. Les auteurs utilisent ces personnages dans des proportions apparemment différentes.

1.3 Émotions

La figure 2 présente la densité de chaque émotion dans les répliques des personnages principaux. Des profils s'en dégagent : Ross a plus facilement tendance à se montrer effrayé par exemple, ou Rachel à être en colère.

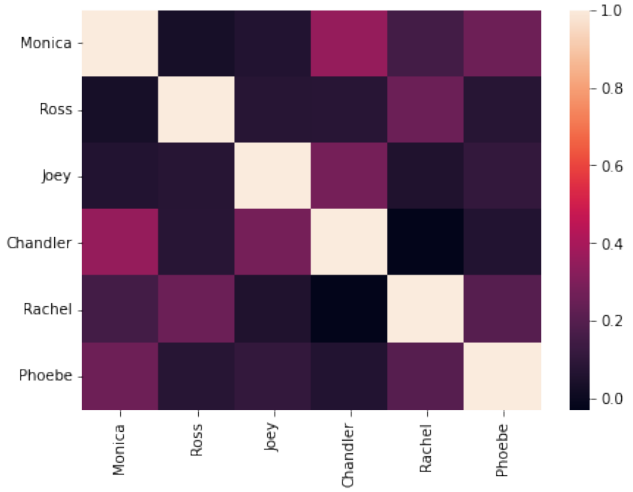


Figure 3 – Corrélation entre apparitions des personnages

1.4 Interactions entre personnages

L'interaction entre les personnages principaux est mesurée comme la corrélation entre leurs apparitions dans les scènes. Cette mesure est reportée en figure 3. Elle met en avant le couple emblématique formé par Chandler et Monica, qui induisent la seule corrélation significative. On constate que le second couple emblématique de la série, formé par Ross et Rachel, ne suit pas la même tendance : les personnages ont plus facilement tendance à apparaître séparément.

1.5 Popularité des personnages

La popularité de chaque personnage est mesurée comme la corrélation linéaire entre la proportion des répliques de l'épisode prononcées par le personnage et la notation IMDb de cet épisode. Cette mesure fournit une seule valeur exploitable : un coefficient de corrélation de 0.185 pour Ross, ce qui en fait le personnage préféré des téléspectacteurs. La représentativité de cet indicateur sera discutée en 2.1.2.

2 Analyse en composantes principales et inertie du jeu de données

2.1 ACP sur les épisodes

Une ACP est réalisée sur le jeu de données comportant les variables quantitatives qui décrivent les épisodes

(nombre de téléspectateurs, notation IMDb et proportion des interventions des personnages principaux), avec un nombre arbitraire de 2 composantes. L'inertie du jeu de données est alors presque entièrement expliquée (à plus de 99%) par le premier axe. Ce phénomène est confirmé par une initialisation alternative de l'objet `PCA` de `scikit-learn` en indiquant non plus le nombre de composantes, mais la proportion de la variance totale à expliquer. Pour une explication de 99% de la variance du modèle, un seul axe est automatiquement choisi.

2.1.1 Influence des téléspectateurs

On s'intéresse ici à la corrélation entre axes principaux et variables initiales. Aucun paramètre particulier n'ayant été passé à l'ACP de `scikit-learn`, on considère n la taille de l'échantillon ($n = 236$ ici) et la matrice des poids $D_p = \frac{1}{n}\mathrm{Id}_n$, et on calcule les corrélations par la formule :

$$\forall \alpha \in [\![1,2]\!], \forall j \in [\![1,8]\!], \mathrm{Cor}(\alpha, j) = \frac{1}{\sigma_j}\frac{1}{\sqrt{\lambda_\alpha}}\mathbf{x}_j^\top D_p \boldsymbol{c}_\alpha.$$

On observe que le premier axe, qui explique la quasi-totalité de la variance, est très fortement corrélé (à plus de 99%) au nombre de téléspectateurs, et légèrement (38%) à la note IMDb. Nous sommes donc en présence d'un effet flagrant du nombre de téléspectateurs sur l'ensemble du jeu de données. Afin d'améliorer la qualité de la représentation, nous normalisons le jeu de données en divisant chaque ligne par le nombre de téléspectateurs, et en réalisant de nouveau une ACP sur les variables restantes. Le même phénomène est observé, avec cette fois-ci la notation IMDb.

2.1.2 Pertinence des conclusions tirées

Ces résultats étaient attendus au regard de ce que nous avons développé en 1.2.1. En effet, nous avons expliqué que les quantités de dialogue de chaque personnage variaient assez peu entre les épisodes. Rien n'est surprenant donc, dans le fait que les axes principaux soient peu corrélés avec les variables donnant les proportions de répliques de chaque personnage.

Si cela va dans le sens de l'hypothèse que nous avons faite plus tôt, nous devons remarquer que les conclusions qui seront tirées du nombre de téléspectateurs et de la notation IMDb des épisodes doivent être replacées dans le contexte de ce jeu de données dont l'inertie est presque entièrement expliquée par ces deux mêmes variables. Ce qui a été affirme en 1.5 est donc à nuancer.

doc_id	paragraph_id	sentence_id	sentence	token_id	token	lemma	upos	xpos	feats
1	1	27	Stop cleansing my aura !	1	Stop	stop	VERB	VB	VerbForm=Inf
1	1	27	Stop cleansing my aura !	2	cleansing	cleanse	VERB	VBG	VerbForm=Ger
1	1	27	Stop cleansing my aura !	4	aura	aura	NOUN	NN	Number=Sing

Figure 4 – Exemple de séparation morpho-syntaxique par le modèle choisi

3 Traitement automatique du langage

Afin d'exploiter les données textuelles à notre disposition sous forme de variables quantitatives, nous avons choisi d'appliquer une méthode de traitement automatique du langage, et plus précisément de *topic modelling*, sur les répliques regroupées par épisode.

3.1 Segmentation et analyse morpho-syntaxique du dataset

Cette phase est celle de pré-traitement des données textuelles.

1. On construit, à l'aide d'un modèle morpho-syntaxique, une table associant à chaque mot des répliques un *lemme* (Forme canonique d'un mot variable) et une classe grammaticale, entre autres.
2. On ne retient que les noms, noms propres, verbes et adjectifs.
3. On exclut de la table des termes trop généraux ou non pertinents pour l'analyse.

On obtient alors un jeu de données nettoyé (*cf* figure 4) qui sert de base à l'analyse.

3.2 Topic Modelling

Le *topic modelling* est une technique de classification à variables latentes des mots d'un corpus de documents. Elle suppose que ces mots sont organisés en thèmes, présents à différents degrés dans l'intégralité du corpus. Son objectif est de déduire, à partir du seul accès aux mots, la structure latente de ces thèmes.

3.2.1 Filtrage des données (*TF-IDF*)

Les lemmes obtenus doivent être filtrés afin de servir de base à l'inférence des *topics* latents. Les lemmes les plus pertinents *pour chaque épisode* sont choisis par un double filtrage, fonction :

— de leur fréquence d'apparition dans l'épisode ;
— de leur pertinence au regard du corpus entier, mesurée par leur valeur TF-IDF.

La valeur TF-IDF (*Term Frequency-Inverse Document Frequency*) d'un lemme est fondée sur la loi de Zipf, qui établit que la fréquence d'apparition $f(\omega)$ d'un mot ω dans un corpus est liée à son rang r_ω dans le classement des fréquences d'apparition par une relation de forme $f(\omega) = \frac{K}{r_\omega}, K \in \mathbb{R}_+$.[8] La mesure TF-IDF permet ainsi de pondérer l'importance d'une unité lexicale (ici un lemme) contenue dans un document (ici un épisode) en fonction de l'entièreté du corpus.

3.2.2 Structure latente des *topics*

Une grande partie dans le processus de recherche des *topics* est de déterminer *a priori* le nombre optimal K^* de topics. Dans cette étude, l'intervalle du nombre K est défini entre 2 et 50. Afin de déterminer K^*, on utilise une fonction faisant intervenir de 4 métriques (CaoJuan[6], Deveaud[5], Griffiths[7] et Arun[4]). Ce nombre doit correspondre aux valeurs les plus faibles possibles des métriques CaoJuan et Arun et en même temps correspondre aux valeurs les plus élevées possibles des métriques Griffiths et Deveaud. Le nombre optimal de *topics* est un compromis entre les résultats proposés par ces différentes métriques. On obtient $K^* = 49$.

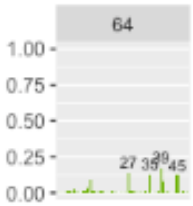
Une fois le nombre K fixé, l'analyse *topic modelling* est lancée afin d'extraire les thèmes en fonction des termes du vocabulaire (ici les lemmes) présentant les probabilités d'apparition les plus élevées dans chacun de ces thèmes.

3.2.3 L'allocation de Dirichlet Latente

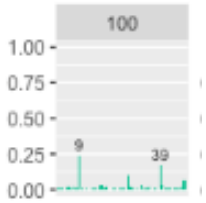
Cette analyse repose sur l'allocation de Dirichlet latente (*Latent Dirichlet Allocation*, LDA). Il s'agit d'un modèle génératif probabiliste à trois couches d'un corpus de documents. Ces couches représentent les échelles de raisonnement : le corpus, le document, et enfin le mot. Chacune d'elles est paramétrée, et l'établissement de lois de probabilités jointes ou conditionnelles permet d'estimer l'un de ces paramètres, qui est celui d'intérêt de notre problème : la probabilité d'appartenance de chaque lemme à chaque *topic*. Ce paramètre correspond à la matrice notée β dans les précisions proposées en annexe B. Précisons cependant que la méthode n'est pas en mesure de nommer les *topics* ainsi définis.

Nous obtenons ainsi une estimation du paramètre β, mais aussi des probabilités d'apparition de chaque *topic* dans chaque épisode. Si de nombreux *topics* sont constitués de termes très généraux, certains font cependant ressortir des thèmes de la série : la figure 5 montre les probabilités d'apparition du *topic* 4 dans des épisodes mettant en scène le mariage de Ross et Emily

Figure 5 – Episodes avec une forte probabilité d'apparition du thème "Londres"



(a) Dispute entre Ross et Rachel



(b) Accouchement de Phoebe

Figure 6 – Apparition des *topics* dans des épisodes marquants

à Londres. Or à ce *topic* sont effectivement rattachés les lemmes "wedding", "London", "Emily" ou encore "Ross". De la même manière, la figure 6a montre que les *topics* 27 et 45, qui concentrent un vocabulaire autour de la dispute, sont présents dans l'épisode 64, qui correspond à la rupture de Ross et Rachel. On constate en 6b la prédominance des *topics* 9 et 39 dans l'épisode 100, qui rassemblent des termes tels que "Frank", "ow" ou encore "hurt", ce qui est cohérent avec le fait que dans cet épisode, Phoebe accouche des triplés qu'elle porte pour son frère Frank.

3.3 Affectation des émotions au jeu de données entier

Puisqu'au sein de la base de données morpho-syntaxique, nous disposons d'un champ intitulé "sentence", nous l'utilisons pour estimer les 8 émotions présentes dans ce corpus : anger, anticipation, disgust, fear, joy, sadness, surprise, trust. Elles sont déterminées à partir du dictionnaire NRC (Mohammad, 2010 [3]) Ces émotions sont classées dans une matrice qui nous servira de base pour la prochaine partie.

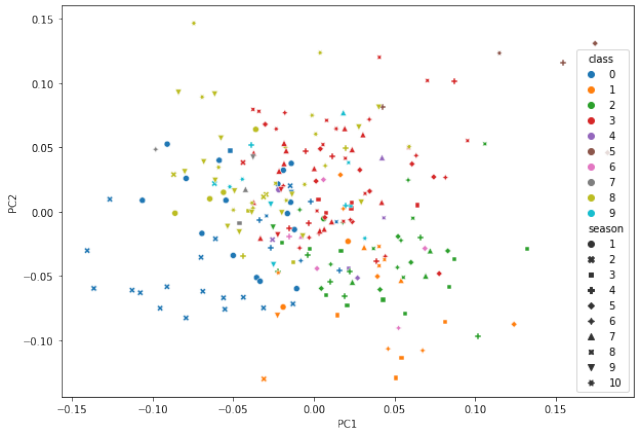


Figure 7 – Classification des épisodes par les *K-means*

4 Classification des épisodes

L'analyse morpho-syntaxique du jeu de données des répliques nous a permis d'obtenir une description quantitative des épisodes par la fréquence de chaque émotion et de chaque *topic*. À partir de ces données, nous cherchons à effectuer une classification des épisodes, afin de voir si une structure logique (une cohérence par saison, par exemple) est identifiable. L'algorithme des *K-means* est donc appliqué aux variables décrivant les probabilités d'apparition de chaque émotion et de chaque *topic* dans les épisodes. La règle du coude motive une classification en 10 classes (ce qui est cohérent avec le nombre de saisons de la série). Le résultat est présenté dans la figure 7.

Cette analyse fait quelque peu ressortir la structure de saison, ce qui permet d'affirmer qu'il existe une certaine homogénéité dans les sujets abordés et les émotions utilisées par saison. Cependant, les saisons de *Friends* sont relativement longues (23.6 épisodes en moyenne) et la cohérence n'est pas toujours assurée entre les épisodes, ce qui explique que cet effet ne soit pas des plus flagrants. On peut néanmoins affirmer qu'il existe en étudiant les saisons majoritaires dans chacune des classes établies. 8 des 10 saisons apparaissent comme majoritaires dans une des classes. Cette majorité n'est jamais flagrante, mais elle nous motive à persister dans la démarche de classification en saisons à partir des données issues de l'analyse morpho-syntaxique.

5 Prédictions sur les saisons

La mise en évidence des structures de saisons par l'algorithme des *K-means* permet d'utiliser cette variable comme variable à expliquer pour des problèmes d'apprentissage supervisé. Soit $\Omega = \{1, ..., 10\}$ l'ensemble des classes auxquelles peuvent appartenir les épisodes, chaque classe représentant une saison. Nous allons ici chercher à construire des classifieurs permettant d'affecter un épisode à sa saison, à partir des valeurs numériques obtenues par l'analyse morpho-syntaxique des répliques.

Nous avons noté plus haut que la scission entre les saisons n'était pas si marquée dans les données obtenues. De plus, le nombre de classes considéré ici est élevé. Pour ces raisons, nous ne nous attendons pas à obtenir d'excellentes performances de nos classifieurs.

5.1 Sélection des classifieurs

Deux modèles peuvent être utilisés ici sans condition :
— les K plus proches voisins ;
— les arbres de décision, méthodes de *bagging* et forêts aléatoires.

Les paramètres optimaux de ce modèles (nombre de voisins ou d'arbres à combiner) sont déterminés grâce à la classe `GridSearchCV` de `scikit-learn`. La pertinence d'autres modèles est à discuter.

5.1.1 Normalité

L'utilisation des modèles d'analyse discriminante nécessite que les données suivent des lois normales multidimensionnelles conditionnellement à leurs classes. Afin de vérifier ceci, nous utilisons le test de Henze-Zirkler pour la normalité multidimensionnelle. Nous fournissons quelques détails à son propos en annexe A, et en utilisons l'implémentation de la librairie `Pingouins`. L'hypothèse de normalité de nos données est rejetée ; nous n'appliquerons donc pas ici d'analyse discriminante.

5.2 Performances

Les performances des classifieurs sont mesurées ici à l'aide d'une validation croisée à 5 plis. Elles sont reportées dans la table 2.

Table 2 – Performances des classifieurs

Classifieur	Score
K-PPV	0.37
Arbre de décision	0.32
Bagging	0.47
Forêt aléatoire	0.51

5.3 Élagage de l'arbre de décision

L'arbre de décision utilisé précédemment n'est pas contraint dans sa phase de construction. Il est donc évidemment sujet au phénomène de sur-apprentissage. Le recours à `GridSearchCV` pour déterminer sa profondeur maximale optimale serait une approche naïve ; nous l'avons donc ajusté en implémentant la méthode d'élagage coût-complexité avec validation croisée. La précision obtenue est à peine meilleure (34%). Les critères de Gini des 2 noeuds de hauteur 1 valent respectivement 0.774 et 0.87, ce qui dénote effectivement une grande impureté du jeu de données.

Les fréquences des émotions apparaissent très peu parmi les variables de décision ; nous avons donc essayé de répéter l'apprentissage avec ces seules variables. Le résultat est sans appel : la procédure d'élagage fournit $\lambda_{opt} \simeq 8 \times 10^{-3}$, ce qui indique que l'optimisation de la complexité doit être sacrifiée pour espérer obtenir une erreur d'apprentissage raisonnable. Pourtant, ce sacrifice n'amène pas au-delà d'un score de 13% (à peine plus qu'une décision aléatoire), ce qui achève de montrer que les émotions sont un très mauvais critère de discrimination des saisons. Un tel phénomène aurait pu être observé par une technique de réduction de données comme l'ACP, et met en avant le caractère linéaire de la série pour ce qui est de la répartition des émotions au fil des saisons.

6 Propos de fin

Le jeu de données qui a retenu notre attention présentait originellement certaines difficultés pour la mise en oeuvre des techniques d'analyse, centrées sur des variables quantitatives. S'il permettait de se pencher sur une analyse descriptive sans encombre, nous avons dû, pour mettre en oeuvre des techniques plus poussées, et notamment des méthodes de classification supervisées et non supervisées, passer par une phase de traitement automatique du langage. Nous avons ainsi pu extraire de nos données textuelles, des grandeurs nous permettant d'observer l'émergence de la structure en saisons. Cette tendance était suffisamment prononcée pour que nous choisissions de la réinvestir dans le cadre de méthodes d'apprentissage supervisé, mais pas assez pour que les résultats obtenus soient réellement significatifs. Remarquons cependant les performances fournies par la forêt aléatoire qui, malgré des données relativement bruitées, et un nombre élevé de classes, atteint un score d'une bonne classification sur deux. Toutes les observations qui ont pu être faites ont été l'occasion de tirer des conclusions, qui ne trahissent que peu l'idée que l'on se fait de la série : assez linéaire, mais dont certaines fulgurances suffisent à marquer.

Annexes

A Test de normalité multidimensionnelle de Henze-Zirkler

Henze et Zirkler proposent en 1990 une famille de tests d'adéquation pour la loi normale multidimensionnelle. [2] Ces tests, pour un échantillon multidimensionnel $n \times p$, reposent sur des statistiques de forme

$$T_{n,\beta} = n\left(4\mathbb{1}_{\{0\}}(\det S_n) + D_{n,\beta}(1 - \mathbb{1}_{\{0\}}(\det S_n))\right),$$

où S_n désigne la matrice de variance-covariance du jeu de données et

$$D_{n,\beta} = \frac{1}{n^2}\sum_{j,k=1}^{n} \exp\left(-\frac{\beta}{2}\|Y_j - Y_k\|^2\right) - 2(1+\beta^2)^{-\frac{p}{2}}\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{\beta^2}{2(1+\beta^2)\|Y_j\|^2}\right) + (1+2\beta^2)^{-\frac{p}{2}}$$

avec $\beta > 0$. Les auteurs montrent que sous l'hypothèse nulle d'adéquation au jeu de données d'une loi normale multivariée, ces statistiques convergent vers des combinaisons linéaires de variables suivant des lois du χ^2. Les coefficients de ces combinaisons sont les valeurs propres d'un opérateur introduit par les chercheurs, que nous ne détaillerons pas ici

B L'allocation de Dirichlet Latente

Les paramètres de la LDA sont proposés par Blei *et al.* ([1]) de la manière suivante :

- α et β sont les paramètres du corpus, où $\beta_{i,j}$ modélise la probabilité que le mot i soit rattaché au *topic* j ;
- θ est le paramètre d'un document tel que $\theta \sim \mathrm{Dir}(\alpha)$;
- z est le paramètre du mot tel que $z \sim \mathcal{M}(\theta)$.

À partir de ces notations, une loi jointe sur tous ces paramètres est proposée par les auteurs.

Références

[1] David M. Blei, Andrew Y. Ng et Michael I. Jordan. "Latent dirichlet allocation". In : *Journal of Machine Learning Research* 3 (2003). DOI : http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

[2] N. Henze et B. Zirkler. "A class of invariant consistent tests for multivariate normality". In : (1990).

[3] Saif M. Mohammad. *NRC Word-Emotion Association Lexicon*. 2010. URL : https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (visité le 25/05/2022).

[4] R. Arun - V. Suresh - C.E. Veni Madhavan - M.N. Narasimha Murthy. *On Finding the Natural Number of Topics with Latent Dirichlet Allocation : Some Observations*. 2009. URL : https://link.springer.com/chapter/10.1007/978-3-642-13657-3_43 (visité le 24/05/2022).

[5] Patrice BELLOT Romain DEVEAUD Éric SANJUAN. *Accurate and effective latent concept modeling for ad hoc information retrieval*. 2014. URL : https://dn.revuesonline.com/article.jsp?articleId=19419 (visité le 24/05/2022).

[6] Juan Cao - Tian Xia - Jintao Li - Yongdong Zhang - Sheng Tanga. *A density-based method for adaptive LDA model selection*. 2009. URL : https://www.sciencedirect.com/science/article/pii/S092523120800372X?via%5C%3Dihub (visité le 24/05/2022).

[7] Mark Steyvers Thomas L. Griffiths. *Finding scientific topics*. 2004. URL : https://www.pnas.org/doi/full/10.1073/pnas.0307752101 (visité le 24/05/2022).

[8] Shuiyuan Yu, Chunshan Xu et Haitao Liu. *Zipf's law in 50 languages : its structural pattern, linguistic interpretation, and cognitive motivation*. 2018. DOI : 10.48550/ARXIV.1807.01855. URL : https://arxiv.org/abs/1807.01855.